DOCUMENT RESUME

ED 398 285 TM 025 631

AUTHOR Crooks, Terry
TITLE Validity Issues in State or National Monitoring of Educational Outcomes.
PUB DATE Apr 96
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Decision Making; *Educational Assessment; Foreign Countries; Models; *National Competency Tests; *Outcomes of Education; Scoring; State Programs; *Testing Programs; Test Use; *Validity
IDENTIFIERS Large Scale Assessment; *Monitoring; National Assessment of Educational Progress; *New Zealand; United Kingdom

ABSTRACT

A recently developed model of validation (T. J. Crooks, M. T. Kane, and A. S. Cohen, 1996) is briefly outlined. It conceptualizes assessment as divided into a chain of eight linked stages: (1) administration; (2) scoring; (3) aggregation; (4) generalization; (5) extrapolation; (6) evaluation; (7) decision; and (8) impact. The model is then used to examine validity issues related to state or national monitoring of educational outcomes. Current procedures for the National Education Monitoring Project in New Zealand are discussed in some depth, with brief comments on validity issues in two other national ass ssment systems, the U.S. National Assessment of Educational Progress and the assessment system of England and Wales. The examples illustrate how assessment strategies are shaped to fit particular interpretations and uses, and how the choices made can limit validity for other purposes. (Contains 1 figure and 20 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
TERRY CROOKS

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 398 285

Validity Issues in State or National Monitoring of Educational Outcomes

Terry Crooks
Educational Assessment Research Unit, University of Otago,
P. O. Box 56, Dunedin, New Zealand

Paper presented at the 1996 Annual Meeting
of the
American Educational Research Association
New York City, April 8-12, 1996

Abstract

A recently developed model of validation (Crooks, Kane & Cohen, 1996) is briefly outlined. It conceptualises assessment as divided into a chain of eight linked stages: administration, scoring, aggregation, generalization, extrapolation, evaluation, decision, and impact. The model is then used to examine validity issues related to state or national monitoring of educational outcomes. Current procedures for national monitoring in New Zealand are discussed in some depth, with brief comments on validity issues in two other national assessment systems. The examples illustrate how assessment strategies are shaped to fit particular interpretations and uses, and how the choices made can limit validity for other purposes.

BEST COPY AVAILABLE

TM025631

Validity is the most important consideration in the use of assessment procedures. The preeminent status it holds is widely acknowledged, in articles, textbooks and professional standards. Recent efforts to build a more coherent and unified view of validity have expanded its scope and further strengthened its importance (Cronbach, 1980, 1988; Kane, 1992; Linn, 1994; Linn, Baker & Dunbar, 1991; Messick, 1989, 1994; Moss, 1992; Shepard, 1993). The breadth and centrality of validity, as now conceived, is clearly evident in Messick's recent definition:

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. (1989, p. 13)

There is, however, a difference between affirming the primacy of validity and acting upon it. As Linn (1994), Gipps (1994) and others have noted, assessment efforts often seem to have given less detailed attention to validity than to reliability or generalizability. This situation has almost certainly arisen because there are better developed and agreed upon operational procedures for examining and optimizing generalizability than there are for examining and optimizing validity. Validity estimation necessarily relies heavily on human judgement, and is therefore very susceptible to various interpretations and emphases. The inclusion of consequences as an important consideration in validity has not made the validation of assessments any easier.

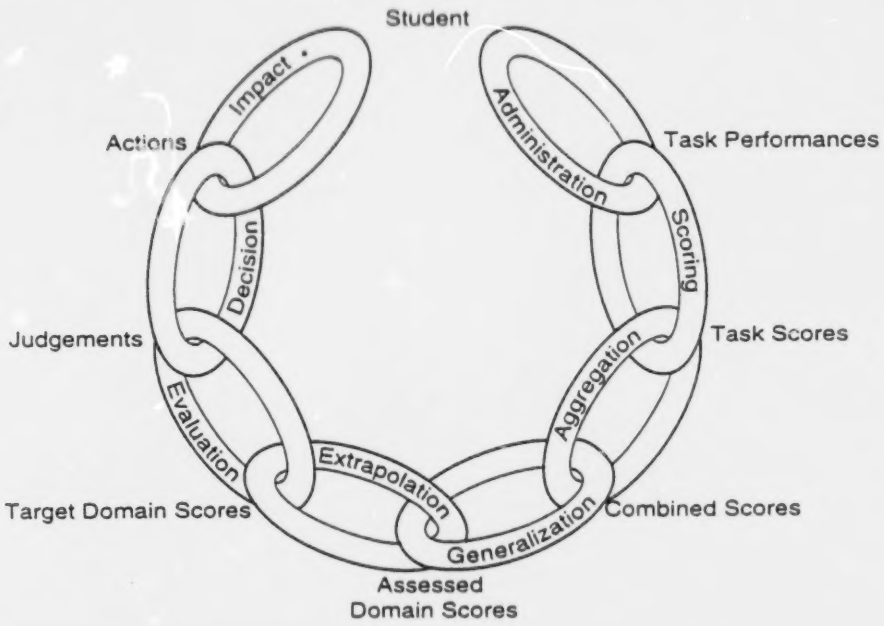
Shepard (1993) has noted that to assist us with the task of validation, we need approaches which help us to organise our thinking about important validation questions and to identify issues which need particularly close scrutiny. One such approach has been to identify sets of validity criteria which should be considered (see, for instance, Cole & Moss, 1989; Frederiksen & Collins, 1989; Haertel, 1985; Linn, Baker & Dunbar, 1991; and Messick, 1995). Another approach has been to build on Cronbach's ideas about validation argument (Cronbach, 1988). This approach has been most highly developed in the work of Kane (1992) and Shepard (1993).

The approach suggested recently by Crooks, Kane and Cohen (1996) aims to combine the virtues of a clearly defined set of validation criteria and the structure of an argument-based approach. Assessment is depicted as divided into eight conceptually distinct stages, with validation then based on careful scrutiny of each of these stages. The eight stages are likened to eight links of a chain, with weakness of any one link weakening the chain as a whole. Further guidance is offered to validators through identification of several examples of threats to validity associated with each of the eight links.

The authors have noted that validation can only take place if the intended purposes of the assessment are well understood. The appropriateness of the assessment tasks and procedures to the purposes must be a central issue in evaluating the strength of each link in the assessment chain. Unintended side effects must be considered alongside evidence that the assessment procedures lead to progress towards the intended purposes.

Crooks et al. (1996) base their model of validation on the model below, which depicts assessment as involving eight linked stages:

1. *Administration* of assessment tasks to the student.
2. *Scoring* of the student's performances on the tasks.
3. *Aggregation* of the scores on individual tasks to produce one or more combined scores (total score or subscale scores).
4. *Generalization* from the particular tasks included in a combined score to the whole domain of similar tasks (the *assessed domain*).
5. *Extrapolation* from the assessed domain to a *target domain* containing all tasks relevant to the proposed interpretation.
6. *Evaluation* of the student's performance, forming judgements.
7. *Decision* on actions to be taken in light of the judgements.
8. *Impact* on the student and other participants arising from the assessment processes, interpretations, and decisions.



The authors illustrated the importance of the eight links by listing an example of the threats to validity associated with each link. They noted that validity may be seriously undermined if one or more of the following circumstances apply: some students receive inappropriate help with the tasks (administration link); scoring of some or all of the tasks emphasises unimportant but easily rated aspects of the performances (scoring link); scores for tasks which are very heterogeneous are added together (aggregation link); few tasks are used, so a small sample of performance is obtained (generalization link); no tasks are included from some substantial sections of the target domain (extrapolation link); performance is interpreted using construct language without supporting evidence (evaluation link); the standards used in making decisions are inappropriately high or low (decision link); or actions resulting from the assessment undermine the educational progress of many of the students (impact link).

Applying the Validation Model to New Zealand's National Education Monitoring Project (NEMP)

Overview of NEMP

After three years of planning and development work (Crooks & Flockton, 1993; Flockton & Crooks, 1994), systematic procedures for national monitoring of educational outcomes in New Zealand commenced in 1995. The main purposes for national monitoring in New Zealand are to identify and report patterns and trends in educational performance, and to provide high quality, detailed information which policy makers, curriculum planners and educators can use to debate and refine educational practices. My co-director and I have a strong commitment to achieving maximum formative value from national monitoring, while not neglecting the provision of information for summative purposes.

Almost all areas of the curriculum are to be monitored on a four year cycle, at two levels in the education system: year 4 (students aged 8 to 9 years) and year 8 (students aged 12 to 13 years). Carefully selected national samples of 480 students (about one percent of the school population at each class level) attempt each assessment task.

The tasks are administered by teachers recruited nationally and released by their schools for a six week period. They receive one week of training and then spend five weeks administering the assessment tasks at schools in their region. A pair of teachers usually spend one week in each sample school, assessing 12 students during that week. Small schools, with less than 12 students at the chosen class level, are paired geographically when schools are sampled.

Students work with an assessing teacher individually or in groups of four students. After an orientation session when the 12 students meet the two teachers and learn about the types of activities they will experience during the week, each student has four assessment sessions lasting up to an hour. In one session, the student works individually with one of the teachers, performing various "hands-on" tasks and being guided and questioned by the teacher. This session is recorded on videotape for later scoring and analysis. In a second session, four students work collaboratively, supervised by a teacher, with the session videotapes for later scoring and analysis. The third and fourth session will vary in nature from year to year. In 1995, the third session had four students working individually around a series of stations in which tasks had been set up (paper-and-pencil tasks and hands-on science tasks), while the fourth session had four students working individually to create two works of art each. Video replay equipment is available in all sessions, so that task instructions can be standardised or enriched through the use of video clips. Randomly chosen subgroups of four students attempt completely different sets of assessment tasks, which means that in total about 12 hours of different assessment tasks are administered in each school, during 21 assessment sessions.

Scoring and analysis is done after the data have been gathered from all students. This work is done at the Project's offices in Dunedin. Tasks which require little

professional judgement in the marking process are marked by senior education students, most of whom are nearing the end of preparation to become teachers. Tasks which require higher levels of professional judgement are marked by teachers, recruited from throughout New Zealand for this work. Most tasks involve several stages, component or aspects, with scoring rubrics for many of the tasks including both analytical and global scoring approaches.

Validity issues related to task administration

Task administration is the first link in the assessment process. If task administration procedures are flawed, reducing the fidelity with which students' capabilities are portrayed through their performances on the assessment tasks, there are no known procedures which can adequately correct for those flaws.

NEMP exists solely to address issues relating to the national system of education in New Zealand. No information about the performance of individual students or the groups of students from particular schools is made available to assessed students, their families, or the sampled schools. Indeed, the light sampling of schools and students and the allocation of three different sets of tasks to students severely limits the possibility of reporting meaningfully on the performance of individual students or schools. As a consequence, the assessments are low stakes assessments.

One threat to validity associated with task administration is that anxiety about the assessment prevents some students from demonstrating their capabilities on the tasks. Given the low stakes nature of NEMP and the care taken in the way tasks are designed and presented, students' anxiety about their performance is not likely to constrain validity. The teachers administering the tasks are instructed to ensure that the students are encouraged to do their best, but also to move students on to other tasks when they are making no progress. Most tasks are structured to allow a confident start, with more challenging sections later in the task.

A second threat involves possible efforts by the students' regular teachers to coach the students for the assessments. The extent to which this can be effective depends on the breadth of the assessment and the availability of detailed information about what will be assessed. Whether teachers will want to try to coach inevitably depends on the stakes involved. The NEMP assessments are very wide ranging and carry low stakes, and therefore are unlikely to be vulnerable to coaching.

A third threat is that students' capabilities are not adequately represented in task responses because the students could not understand the tasks or could not communicate their response adequately. For instance, a student's performance on a science task may be undermined by their inability to read the task instructions or to write about what they have done and understood. In NEMP, with a maximum of four students working with a teacher, students can receive instructions orally or can ask for help with writing their response. Furthermore, half of the assessment is video recorded, with students mainly responding orally rather than in writing. Some tasks are presented partly or fully on video, providing visual and oral instructions for the students. These factors help students to understand tasks and communicate

their responses. A further feature of NEMP has the potential to either enhance or undermine validity. The teachers are encouraged, if they think a child does not understand a task, to repeat or rephrase the instructions or to use additional prompts. At best, this helps the students understand what is required of them, at worst it provides undesirable cues about how to respond to the tasks.

A fourth threat associated with task administration poses a greater threat to the validity of NEMP assessments. Because of the low stakes nature of the assessments, there is a serious risk that students will not be sufficiently motivated to do well on the tasks, and therefore will fail to demonstrate their true capabilities. Student motivation has been a concern with NAEP assessments. Our response to this threat in New Zealand has been to place considerable emphasis on making tasks seem interesting and worthwhile to students. We try to embed tasks in contexts which most students will find relevant to them, to include elements of fun or adventure where possible, and to ensure that the teachers present the tasks in ways that make them palatable for the students.

A final issue, peculiar to situations in which assessment information is aggregated across individuals, is the extent to which the sample of students who responded to tasks is representative of the population to which generalization is to take place. If representative sampling is not used, or if selected schools or students refuse to participate, or if data is lost through absentees, incorrect administration or equipment failure, generalization to the population is undermined. For the first year of NEMP, these factors are not likely to pose major risks. The sample was a carefully selected random national sample, only 2 of 240 schools originally selected did not participate, less than 2 percent of students were excluded at the request of their schools or parents, and usable data was obtained on most tasks from about 95 percent of the remaining sample of students.

Validity issues related to scoring

When scoring assessment tasks, the aim must be to capture in the scores the richness of performances on the task. The scoring should give credit for the most important qualities associated with the task, avoid biases, achieve a good level of consistency (within and between markers), and provide a sufficiently detailed analysis of performance for the intended purposes of the assessment (such as feedback on strengths and weaknesses).

Scoring is a major challenge for NEMP. None of our tasks are machine scored, and about 80 percent of the tasks require non-trivial judgements to be made. Each year, about 3000 hours of the task performances are recorded on video, and most of these tasks can only be scored by viewing the video (some of the video recorded tasks also have paper-and-pencil components). Other tasks involve substantial student performances, such as carrying out scientific experiments or creating works of art. Scoring involves thousands of hours of work by more than 100 teachers, and further thousand of hours of work by senior university students. For all these reasons, scoring must be viewed as a probable weak link in the validity of NEMP assessments.

We have adopted several strategies to control these threats to validity. Tasks requiring high levels of professional judgement are scored by experienced teachers, selected because of their perceived expertise in judging student work in the class levels and curriculum areas involved. Some of the most demanding tasks are scored by pairs of teachers, so that a consensus process moderates the perceptions and judgements of individual scorers. Selected examples of student performances are used in training the scoring teams, and interpretation of the scoring rubrics is discussed in advance and after each scorer has marked several students' work. About ten percent of the scoring of each scorer or scorer pair is check-scored by other scorers, so that inter-rater consistency can be examined. Where necessary, the check scoring information, together with descriptive statistics on the distribution of scores for each scorer, can be used to design scaling algorithms and adjust the scores awarded by scorers who appear to have been too lenient or too demanding.

The scoring rubrics are unique to each task, reflecting the intention that performance on each tasks will be reported separately. Nevertheless, common approaches are used wherever they seem appropriate, helping scorers to adapt more quickly to new tasks. Most tasks include a mixture of analytic and global scoring approaches: scorers are asked to record particular features present in students' responses, to rate various aspects of the responses on defined scales, and to sum up with a global rating covering several aspects of the students' performance. This combined focus on both specific features and the overall performance should allow us to give reasonably detailed reports on the strengths and weaknesses of students' work, as required for the formative function of our national monitoring. It will be interesting to study the relationship between more specific attributes and the global ratings, because this will give information about the relative importance judges assigned to the specific attributes.

Validity issues related to aggregation

Aggregation is the combining of information from several tasks to derive scale scores. If the tasks combined are too diverse (with low correlations between tasks), the scale scores may be difficult to interpret, and have low generalizability. Furthermore, unless the inter-task correlations are quite high, the particular mix of tasks will have a strong influence on the scale score, with the threat that different aspects of performance may be inappropriately weighted in the scale as a whole.

Several factors were involved in the decision that reports on student performance in NEMP would focus on individual assessment tasks, or small clusters of highly related tasks.

First, the curriculum in New Zealand schools is undergoing major revision. As a consequence, it would be very difficult to reach agreement on the mix of tasks to be included in highly aggregated scales, such as scores on "science" or even "the living world". Extensive processes of consensus building would be required, and this was incompatible with the limited budget and short development period available to NEMP. Furthermore, the relative emphases in the scales used in 1995 might not be appropriate for use in 1999, the next time the same areas were to be assessed.

Second, a major function of NEMP is formative: to promote debate and action about curricula and educational practices. Reports which focus on highly aggregated scale scores are of little value for this purpose. Indeed, they tend to focus attention on summative issues.

Third, NEMP is using a high proportion of tasks which involve hands-on performance, teamwork, or extended answers. The preponderance of recent evidence suggests that such tasks are likely to have low inter-task correlations, the very condition which makes aggregation most risky (Linn, 1994).

Fourth, NEMP is using multiple-matrix sampling and trying to distribute similar tasks across three sets of tasks, attempted by three non-overlapping samples of students. This would hinder the development and analysis of scale scores.

The decision to use little or no aggregation has important consequences for all subsequent links of the assessment chain. Those consequences will be examined during discussion of those links.

Validity issues related to generalization

Generalization is the process of interpreting scores or performances based on a particular sample of individuals, assessment conditions, occasions, and tasks as indicative of the performance which would have been observed with different samples drawn from the same population of these facets. The extent to which such generalization is justified can be described by estimated variance components and summarised in generalizability coefficients.

In NEMP, the student and occasions facets are not likely to significantly limit generalization of the national data. The reported results use performance data aggregated across 480 randomly sampled students (from 120 schools), assessed in varied circumstances across a five week period, and with low refusal and attrition rates.

The diversity of task formats and the very low levels of aggregation mean that the usual generalizability benefits from aggregating across highly correlated tasks will not apply to NEMP. In effect, each task must stand on its own. Comparisons across a four year time span will use the same tasks, and will attempt to duplicate the assessment conditions as closely as possible. Because of the experience of NAEP with the so-called reading anomaly in the 1986 assessments, we will try to retain selected complete sets of tasks to use four years later, so that the contexts surrounding individual tasks are as consistent as possible.

It could be argued that generalizability across tasks is not an issue, because results will be presented for individual tasks. Interpretations of performance on the tasks will, however, almost certainly involve some comment about the skills demonstrated and the task aspects which were handled well, with the implication that students would have performed similarly on similar tasks. Thus the issue of generalization across tasks cannot be avoided. There is ample research evidence that apparently very

minor changes in task instructions and formats can lead to major differences in student performance (see, for instance, reports on the APU Mathematics assessments in England), indicating that generalized interpretations of performance on individual tasks are potentially troublesome.

Validity issues related to extrapolation

Extrapolation involves making inferences from student performance in the domain of tasks, conditions, and occasions which formed the sampling domain for the assessment (the assessed domain) to student performance in the domain which is used for interpreting the assessments (the target domain).

NEMP staff work collaboratively with a national advisory committee in each curriculum area to develop a concise framework showing the intended outcomes for that curriculum area. This framework is then used to guide the development and selection of assessment tasks. Repeated check are made to ensure that major curriculum strands and emphases were adequately represented in the assessment tasks, and that important knowledge, skills and attitudes are tapped. It is not necessary to worry about the precise balance of tasks, because task results are to be reported separately rather than in aggregated form (more extensive consensus procedures are needed in NAEP, where results are reported in highly aggregated ways).

In many assessment programmes, the range of tasks is significantly constrained because of practical considerations such as the use of whole classes of students, administration of the tasks by the regular class teacher, and limited budgets for scoring. In NEMP, we are in the fortunate position of having few practical constraints on our tasks. The small sample size allows us to spend substantial sums assessing each student. We have no more than four students working with each teacher at any one time, the teachers have a week of specialised training before administering the tasks, video equipment is available to present tasks or task information and to record responses, and a wide range of equipment and supplies can be provided. Our main constraint is time: no task is permitted to take more than an hour, because we do not want tasks spread over more than one session. This limits the magnitude of projects which students can undertake, compared to what they can do over some days in the classroom. For instance, when we ask students to create a painting, they may not have sufficient time to complete it to the level of detail they would like, or to experiment with possible ideas and then prepare a final version. Occasionally, equipment is an important constraint. In 1998, when we are to assess performance in physical education, we will not be able to assess swimming because many of the schools will not have a suitable swimming pool in their grounds.

Despite all of these favourable factors, it is a constant challenge to ensure that we are assessing what really matters. It is by no means easy to develop good tasks to assess some of the more sophisticated outcome we are trying to assess, and even quite small defects in tasks, their administration or their scoring can mean that information about important outcomes is lost.

Validity issues related to the evaluation and decision links

In the model of assessment I am using, evaluation is the process of making judgements on the basis of assessment information. These judgements lead in turn to decisions on actions to be taken as a result of the assessment information. I will discuss these two aspects of the model together.

We have not yet reached the stage where reports on the first year of assessments have been produced and distributed. These reports are intended to be factual reports, giving information about how the students have performed. Most of the judgements and decisions will be made by other stakeholders, after they have received and analysed the reports. Nevertheless, NEMP staff have three major responsibilities which will impact significantly on the validity of the judgements and decisions which arise from the reports.

First, we must try to ensure that the information is presented clearly and in ways which suit the capabilities and needs of different stakeholders. A detailed report which would prove fascinating to many curriculum specialists might be too detailed and hard to follow for many classroom teachers and parents. Accordingly, we are planning to prepare two or three versions of the reports on each curriculum area, tailored to suit different audiences, together with a technical report which provides organisational detail on the annual programme of assessment.

Second, we must be careful in the language we use to interpret particular findings. As Cronbach, Messick and others have noted, it is almost impossible to use describe student performances in words without using constructs. Nevertheless, we need to recognise that some constructs involve much more inference than others. To say that students had little success in solving a particular problem involves little inference, to say that they are not good at solving problems of that type involves more inference, and to say that they showed themselves to be "weak in problem solving skills" involves a high level of inference. We should avoid high inference interpretations in our reports.

Third, we should help structure the discussion that will follow release of the reports. One possible approach is to arrange two or three pre-release discussions of the reports by selected groups of stakeholders (curriculum experts, educational administrators, leaders of teachers' organisations, classroom teachers, parents, and other community representatives). Summaries of the main points agreed in those discussion could be publicly released simultaneously with the reports, and members of the discussion groups would be available to brief their colleagues and talk to the media about the reports.

Validity issues related to impact

Consideration of the consequences of an assessment activity is, in my view, a necessary aspect of the validation of the assessment activity. Without evidence that the assessment serves a useful purpose and does not cause undue harm or hardship, the merits of the assessment activity cannot be properly considered.

NEMP has been carefully designed to minimise negative effects for participants and the education system it monitors. The use of small samples of schools and students, together with the NEMP policy that results of individual students or schools will not be reported, prevents the assessments becoming high stakes activities. These features minimise the risk that schools, teachers or students will be anxious about the assessments. Considerable care is taken in the organisation of the assessments to ensure that participants are well informed about what is involved and that they have the right to refuse to participate or to withdraw at any time. The success of the measures is indicated by the fact that in 1995, when 256 schools and 2870 students were included in a ten week period of assessment, only one telephone call was received from a school or a parent raising any concern about the process (note that an 800 number was available and widely publicised, to reduce obstacles to such phone calls).

Several areas of positive impact are already apparent. We have considerable evidence that the vast majority of students enjoyed being involved in the assessments, and found them intellectually stimulating. The students rated their enjoyment of each task (or cluster of small tasks) they experienced, and we received many comments from parents and the students' regular teachers stating that the students were keen to attend the assessment sessions and came away excited about their experiences. Many students were introduced to content or skills that they had previously had little exposure to, and were able to gain knowledge and skills through their participation.

The 96 teachers who administered the tasks in 1995 were similarly positive about their experiences. They reported that the experience of working so intensively with individual students and with a wide range of interesting tasks gave them insights which would be very useful when they returned to their own schools. Several of the 1995 teachers reported that they had made significant changes in their teaching of the areas they had assessed: one teacher working in special education decided that she should start teaching science to her class because she could now see that there were worthwhile science tasks which her students could undertake and learn from. A further benefit was the opportunity to spend an extended period in each of several different schools, and to pick up good ideas from the teachers in those schools. The 96 teachers involved in 1995 represent more than 0.5 percent of all primary and intermediate school teachers in New Zealand. Our intention is to appoint different teachers each year, so that a steadily increasing proportion of teachers will have experienced involvement in NEMP.

About 150 teachers have been involved in the marking of the assessment tasks used in 1995. Most of them spent 20 hours on marking during a one week period, but some were employed for two or three weeks. Again, they reported that the experience was valuable and would have substantial application in their teaching. Several art teaching specialists, for example, decided that they could substantially improve their judgements of students' work through the use of a structured marking rubric.

Beyond these benefits for participants, we have yet to see what impact our reports will have on the development of educational policies and practices, and on public perceptions of the performance of our education system. I believe we have

established approaches and conditions which will produce substantially greater benefits than negative side-effects, but it will be some years before that belief can be confirmed.

Some Observations About Validity Issues in Other State or National Assessment Procedures

In this brief section I will identify some important validity issues associated with two other national assessment systems. Because my comments are brief and focused mainly on areas of concern, they should in no way be construed as a comprehensive or adequate account of validity issues for the systems discussed.

NAEP

The National Assessment of Educational Progress (NAEP) has a long and distinguished history. Starting in the 1960s under the leadership of Ralph Tyler, it has developed and adapted in quite major ways over the following 30 years. I will restrict my comments to four features.

In its early years, NAEP used a wide variety of task formats, including hands-on performance tasks and tasks undertaken by groups of students. In following years, budget constraints and practical difficulties undermined the initial idealism, and paper-and-pencil tasks became the norm. In the past few years, however, there has been a resurgence in the use of hands-on tasks, with the use of multiple-choice items declining proportionately. This resurgence appears to allow a more valid sampling of student learning outcomes, especially in disciplines involving a large practical component (such as science, art, and music) but also in other areas with expanded curriculum emphases (such as mathematics and literacy).

In its early years, NAEP reports were based predominantly on individual tasks. More recently, they have been based predominantly on aggregated performance across whole curriculum areas or major strands within them. The change occurred because the early reports appeared to create little public interest. As a consequence of the change, however, the potential formative value of the monitoring declined considerably, with correspondingly greater emphasis being given to public accountability (the "Nation's Report Card"). It could be argued that the change increased validity for accountability purposes, but at the expense of validity for formative purposes. Perhaps the sheer size of the US education system makes the formative use of NAEP very difficult to achieve, however the results are aggregated and reported.

In the last eight years, NAEP has begun to report achievement levels for individual states, as well as for the United States as a whole. This has two important consequences. First, it raises the stakes of the assessment significantly, with much publicity being given to differences in performance levels between the states. This

increases the risk that states or schools within them will feel pressure to manipulate the data, by such strategies as excluding a higher proportion of students with disabilities from participants. This, in turn, could undermine the validity of NAEP for its initial purpose of presenting a valid national picture. A second consequence of reporting state performance levels has been an increase in the required sample size, with substantial samples now required in each state. The increased sample size results in increased cost and complexity, with a corresponding pressure to adopt more economical approaches for assessment and scoring. The current popularity of performance assessment has prevented such cost cutting, with the result that cost escalation is a significant threat to NAEP.

My final point is to note that the greater use of hands-on performance tasks may become increasingly incompatible with the current levels of aggregation and generalizability in reporting of NAEP results. More diverse tasks, in a wider range of presentation and response formats, can be expected to have lower average intercorrelations, and it may be appropriate to consider clustering them into smaller subsets. Lower levels of scale generalizability will apparently need to be accepted, whether or not smaller subsets are used.

National Assessment in England and Wales

Over an eleven year period from 1978 to 1988, the Assessment of Performance Unit reported on student achievement in England, Wales and Northern Ireland (Foxman, Hutchinson and Bloomfield, 1991). The APU used carefully-drawn national samples of students, and rich and intensive assessment approaches which allowed considerable attention to the processes which students used as well as the outcomes they achieved. The goal of the APU assessments was to present national data on student achievement and attitudes. Schools and Local Education Agencies had considerable freedom to plan curricula and assessment procedures for local purposes.

In the late 1980s, however, the British government moved forcefully to change this situation. National curricula were developed for England and Wales, and national procedures for assessment were designed and implemented. A central feature of the new system was that every student would be assessed on a similar basis, and the results of assessments at several age levels would be used for three purposes: reporting to parents, providing public information about individual schools, and national monitoring of educational outcomes. The APU was disestablished.

Although the government's initial intention appeared to be to require national paper-and-pencil tests, a national task force developed approaches which appeared to offer a more satisfactory and broader approach (Department of Education and Science, 1988). It suggested the use of teacher assessment focused on the strands of the national curriculum statements, moderated by the use of nationally set tasks. These tasks would be used by the teachers, within their own classrooms and within the context of normal learning activities. In the event, however, the government required prime emphasis in reporting on the results of nationally set tasks, and the pressures on teachers resulted in movement away from extensive use of hands-on performance tasks and towards reliance on paper-and-pencil tests (Black, 1994).

These recent experiences in Britain illustrate several points about the validity of system-level assessment. Most importantly, they demonstrate that attempting to use the same assessment for multiple purposes involves major compromise. No one set of assessments can be optimal for reporting to parents, reporting on schools, and reporting on national patterns of achievement.

There is little doubt that the APU approach to system monitoring was capable of presenting a more valid picture of national achievement patterns than the new system can manage. The huge number of students involved in the current assessments has forced the abandonment of the more interactive task formats used in the APU surveys. Multiple-matrix sampling in the APU survey allowed many more tasks to be used than can be accommodated in the national tests. The high stakes nature of the assessments has increased the threat of debilitating student anxiety and increased the likelihood that teachers will try to teach to the tests.

Another important concern about the validity of the new system relates to its impact on participants in the education system. There is little doubt that the assessments have helped to enforce implementation of the new national curriculum statements, and that in some respects this has been beneficial. However, externally imposed requirements often threaten local initiative. In this case, teachers have had their freedom to respond to perceived local needs constrained, and the extent to which their professional judgement about their students' achievements can be trusted has been questioned. These effects are not likely to be helpful.

Conclusion

This model for considering the validity of educational assessments which was outlined in the first section of this paper has been shown to be useful for analysing the validity of procedures for monitoring educational outcomes in New Zealand and elsewhere. The systematic approach suggested by the model helps to ensure consideration of a wide range of possible threats to validity, from those associated with task development and administration to those associated with the impact of assessments results and decisions on participants. Optimal design of an assessment system requires careful consideration of the intended application of the assessments, and the implications of the application for validation. The use of one set of assessments for multiple purposes almost certainly cannot be optimal for more than one of those purposes.

References

Black, P. J. (1994) Performance assessment and accountability: the experience in England and Wales. *Educational Evaluation and Policy analysis, 16,* 191-203.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. In *Proceedings of the 1979 ETS Invitational Conference* (pp. 99-108). San Francisco: Jossey Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Crooks, T. & Flockton, L. (1993). *The design and implementation of national monitoring of educational outcomes in New Zealand primary schools.* Dunedin, New Zealand: Higher Education Development Centre, University of Otago.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. Unpublished manuscript.

Department of Education and Science. (1988). *National curriculum: Task Group on Assessment and Testing: A report.* London, Department of Education and Science and Welsh Office.

Flockton, L. & Crooks, T. (1994). *Assessment tasks for national monitoring of educational outcomes in New Zealand primary schools.* Dunedin, New Zealand: Higher Education Development Centre, University of Otago.

Foxman, D., Hutchinson, D. & Bloomfield, B. (1991). *The APU experience 1977-1990.* London: School Examinations and Assessment Council.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Gipps, C. V. (1994). *Beyond testing: Toward a theory of educational assessment.* London: Falmer Press.

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research, 55,* 23-46.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23* (9), 4-14.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), 15-21.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23* (2), 13-23.

Messick, S. (1995). Standards of validity and validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Moss, P. A. (1992). Shifting conceptions of validity in educational assessment: Implications for performance assessment. *Review of Educational Research, 62,* 229-258.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405-450.